A New Tool for Prioritizing Candidate Genes

Ayyoub Salmi*, Romain Philippe**
Veronique Blanquet** Ahmed Moussa*

*Systems and Data Engineering Team
University Abdelmalek Essaadi, Tangier, Morocco
**Animal Molecular Genetics Unit
University of Limoges, Limoges, France

Abstract. The identification of genes implicated in a disease or in a complex trait from high-throughput experimental studies remains a challenge. Gene prioritization methods play an important role in identifying the most prominent genes in a study. In this paper we present a new tool for prioritizing candidate genes by applying text mining to data related to genes. We tested our tool using genes near a Quantitative Trait Locus (QTL) related to bone weight in cattle, and differentially expressed genes related to Parkinson Disease.

1 Introduction

The recent advances in sequencing technologies led to an increase in organisms with a completely sequenced and, more importantly, annotated genome Reuter et al. (2015). However, the task of identifying genes that are involved in a particular disease or related to a trait remains challenging.

Candidate gene studies evaluate genetic variation in the region of genes that are physiologically suggested (candidates) to be involved in disease pathogenesis Holloway and Prescott (2017) or related to a desired trait.

In the present article, we describe a new ranking tool that uses text mining to prioritize candidate genes following some criteria using gene expression, gene function, homology, knock-out studies and a survey of literature.

2 Methods

In order to use the text mining approach, we first collated gene related information from different sources. For this purpose we created a database containing 56813 human genes including coding genes, non coding genes and pseudogenes. We used Elasticsearch, a distributed full-text search engine to store and interrogate those documents.

Since we are using a text mining approach, a set of keywords describing the the disease or phenotype under study must be provided. The drawbacks of using keywords